Submission - AI Accountability Policy Request for Comment

A Notice by the National Telecommunications and Information Administration on 04/13/2023

Author: Kassandra Popper - Software Developer

Date: 06/12/2023

(Sections are titled and numbered with the section header and number of the question in the guidance document of the RFC to which they relate, respectively)

AI Accountability Objectives

(1.)

AI Accountability mechanisms should primarily be concerned with ensuring AI products function as advertised, and to a lesser extent according to the norms for the industry and market the product is intended for. As such, the appropriate accountability mechanisms will be dependent on the stated application.

Internal audits are most useful for ensuring product quality is assured on a regular basis, when the incentives of producers and consumers are most tightly aligned.

External audits are most useful for investigating demonstrated problems with a product, to get to the bottom of a major problem which individuals within an organization might have incentive to not be honest about or even cover up/fake evidence for a failure.

Because of necessary application specific implementation details of Trustworthy AI goals, it may make sense to fold AI audits or assessments into accountability mechanisms that focus on other goals such as human rights, privacy protection, security, and diversity, equity, inclusion and access.

AI Accountability practices could be meaningful even in the absence of legal standards and enforceable risk thresholds because members of the nascent AI industry will have common incentive to demonstrate that their systems meet the goals of Trustworthy AI. If AI Accountability practices are designed well, compliance with them can be an advertisable competitive advantage for AI industry participants. Courts, legislatures and rulemaking bodies have a role in sketching out the shape of good AI accountability practices, but to intercede in the business practices of AI industry participants only if failure to follow good practices demonstrably results in harm to the public.

(2.)

The value of certifications, audits, and assessments is mostly to promote trust for external stakeholders, and should not be overly concerned with internal processes at AI organizations. Attempting to regulate and specify the internal processes of firms in a fast evolving technical industry like AI is bound to result in attempts to match behavior to an outdated specification with questionable benefit with respect to the goals of Trustworthy AI. The point where trust is relevant is when the AI product leaves the firm, so policy should be designed to assess that the product leaving the firm is safe, and not concern itself with how it is produced, which is largely irrelevant to its impacts on society. The only exception to this is that the AI firm's internal production processes should be ethical, but that is unrelated to the goals of Trustworthy AI as defined in this RFC's guidance.

(3)

There exist tradeoffs between the goal of an AI system to be effective and other goals related to limiting outputs within certain thematic boundaries and definitions. For example, an AI system cannot effectively provide information about propaganda and misinformation without also providing examples of such potentially harmful material. It is the nature of AI systems that they to some significant degree decide their own outputs, and attempting to limit that flexibility for the sake of other system objectives may cause significant disabling of AI system effectiveness. Measures to satisfy goals such as to not substantially contribute to harmful misinformation or protecting privacy may negatively impact the quality, utility and reliability of an AI system's outputs. In general, requirements for such goals should only be placed on systems where the intended application requires it, and compliance assessments and actions should take into account that to a large degree the production of undesirable outputs is an unsolved problem in generative AI c.f. The hallucination problem in Large Language Models.

Regarding other potential AI accountability goals such as AI systems not substantially contributing to harmful discrimination against people, being safe and legal, these issues are best treated within the context of the particular application domain of an AI system's deployment, and it may be difficult, or make little sense, to try to apply general rules to how AI systems, in general, should treat these broad thematic categories. That AI systems do not harm, discriminate, act unsafely or act illegally is clearly desirable, but the specifics of particular AI system requirements will default to what's already been established as acceptable and legal behavior where they are deployed.

The correct degree of adequate human alternatives for AI systems can probably be left to be set by market mechanisms, as companies will have an incentive to balance system cost with customer satisfaction. Similarly, adequate transparency and explanation to affected people about the uses, capabilities, and limitations of the AI system will likely be set naturally by customers selecting to do business with firms providing products and services with the most satisfactory levels of transparency and explanation of the AI system.

(4.)

AI accountability mechanisms can effectively deal with systemic and/or collective risks of harm by being structured so as to incentivize that AI industry participants strive to distribute the benefits of AI widely and efficiently. By spreading the benefits of new AI technology to the most people who can benefit from and innovate on top of them, any emergent risk will have the broadest and most diverse pool of powerfully AI enabled solvers to counter it. The real risk is in not spreading the benefits of AI widely.

(7.)

Over-regulation of a nascent industry is a peril to be avoided, especially when that industry has the economically powerful promise which AI seems to possess. Too onerous accountability mechanisms could seriously impede the progress in AI R&D in both industry and academia, which would counter-intuitively delay the arrival of more trustworthy AI systems. Progress in software development has greatly benefited from the freedom with which software engineers and firms have to pursue new solutions, without having to apply for special licensing or setup special compliance infrastructure beforehand. Accountability mechanisms which add financial costs to participation in AI software development, even as small as $1000 or less, would likely eliminate many if not most of the potential contributors to this promising field. Much of the current boom in AI application research is building on top of open source AI projects based on github, and many such projects would never have started if the developers or users were required to apply or pay for a license before they could get started.

Additionally, many AI software projects and research studies are using open source AI models, or ones with fairly liberal licensing, such as Meta Inc.'s LLaMA large language model, or the open source image generation model Stable Diffusion from Stability AI. AI Accountability mechanisms should not impede companies from releasing models such as these. The empirical evidence says that these have produced far more good in terms of utility for customers and potential to spur development of useful products than any of the proposed harms having come to pass.

Accountability Subjects

(15.)

AI Accountability efforts should focus on two parts of the value chain: 1.) the collection of data for use in training a machine learning system, and 2.) the distribution of the AI system built from that machine learning system to the customer, citizen or other effected person.

To account for the variable context of downstream deployment of an AI system, accountability mechanisms should typically focus on the point of final application, rather than on the capabilities of frontier models before they have been fine-tuned and packaged for use by customers. If an objective is to create a frontier model that has unbiased behavior, focusing on the point of data collection may be sensible, but it won't account for biases or correction

processes which may be possible in the data processing, so even for this goal a focus on final product is likely to be the most reliable.

(16.)

AI Accountability mechanisms should focus on the application of an AI system, because that is the point at which they impact society. Machine learning by its very nature involves abstruse methods for the production of artificially intelligent artifacts, more so than typical software development, and attempting to standardize these quickly evolving processes will be highly prone to dysfunction and possible unintended consequences for AI industry evolution. The proper place of focus for AI accountability is the point it is being productized and sold to a customer as fit for a purpose. Attempting to regulate “frontier models” whose definition is likely to change with the next machine learning paper uploaded to the online computer science preprint server arxiv will only slow progress in this vital field, blunt economic benefits from AI, and with no benefit to the goals of Trustworthy AI.

(17.)

AI Accountability measures should be mandatory for those applications where there would otherwise be a significant risk of serious or permanent harm to people or property being caused by the AI system in its normal deployment context.

(18.)

In general AI systems should not require quality assurance certifications to be released because such assurances aren’t necessary for the vast majority of likely AI applications. As current state of the art AI products become incorporated into existing software products, AI will soon include most software applications. Adding such a requirement that the AI needs to be certified would have a chilling effect on AI startups and software startups generally, without any apparent benefit, because its not clear what general, non-deployment specific qualities, need being assured for the goals of Trustworthy AI, which aren’t already well handled by existing market and legal constraints. Applications which involve higher risk might reasonably benefit from some quality assurance certification process, but it again comes down to the issue of what qualities need assurance, and that will depend on the application, much less so on the generic properties of the AI system or frontier model that was used as components of the total system.

(19.)

The government will have special needs that will only partly overlap with the needs of private sector AI applications, so it's reasonable to expect that public sector projects may have a mix of common and distinct AI accountability measures. It will again depend on the particular application for what makes sense, and may sensibly be delegated to the individual department or industry oversight committee for determination of appropriate accountability mechanisms.

Barriers to Effective Accountability

(26.)

The lack of a federal law focused on AI systems is hardly a barrier to effective AI accountability. AI systems are ultimately software applications, and governments at all levels, from municipal to state to federal and international levels, have many legal and regulatory tools available to address the impacts and accountability of AI software systems and their producers, as much as with any software. In general, leaving the specification of AI accountability mechanisms to the most granular governmental level at which its effects manifest seems preferable for reasons of customization of the mechanism to the specific conditions of its application domain. It's not clear that a new federal law is presently required for effective AI accountability.

(28.)

In order to promote a robust AI industry which efficiently develops and distributes machine learning innovations and their benefits, accountability mechanisms must not place a burden on new entrants and should minimize the friction required to accountably publish AI research artifacts and new products to market. The best way to consider costs in relation to benefits is to consider the ubiquity of open source software powering today's tech behemoths and AI startups both, and how development of the tools powering the nascent AI boom happened because anyone could publish an AI package for pytorch (an open source AI framework developed by Meta AI) without needing a license or certification. Whereas the costs for this freedom of publication and sharing of methods has been insubstantial. Any imposition on individual or small groups of software developers, open source projects or new AI firms should need extremely strong justification because the health of the AI software and product ecosystem depends on their continued freedom to innovate.

(29.)

Useful operationalizable measures for assessing reliability and trustworthiness of AI systems would benefit the AI industry and customers by providing useful statistics on which to compete for improvements that would be relevant to both customers and industry observers. "AI Alignment" has become malleable in meaning and is used as both an unverifiable standard for hypothetical super-intelligent systems while simultaneously being co-opted into a hollow marketing term. Better measures are needed

AI Accountability Policies

(30.)

AI accountability policy should be sectoral, because each sector will have its own application requirements which are unlikely to overlap significantly enough to justify a harmonized accountability mechanism.

AI Accountability regulation should focus on inputs to validation, in particular on data required to assess that a product is functioning as advertised and is fit for its intended purpose. It should not require increased access to AI systems for researchers or auditors except as necessary for investigating some demonstrated failure of an AI system product that is worthy of an in-depth investigation, such as in cases of serious or permanent harm caused by the AI system during normal operation. Accounting measures should not be mandated unless there is otherwise a significant risk of serious or permanent harm resulting from the AI system's normal operation.